Efficient Uplink Transmission in Ultra-Dense LEO Satellite Networks With Multiband Antennas

Tianjiao Chen, Jiang Liu, Qiang Ye, *Member, IEEE*, Qinqin Tang, Weiting Zhang, *Member, IEEE*, Tao Huang, *Senior Member, IEEE*, and Yunjie Liu

***Abstract*— In this letter, we propose an efficient uplink transmission scheme over ultra-dense low earth orbit (LEO) satellite networks with multiband antennas. A two-timescale transmission resource allocation and frequency band switching optimization framework is established with the objective of maximizing the overall data rate while reducing the band switching cost. The small-timescale transmission resource allocation problem is described as a Markov decision process (MDP) considering continuous LEO satellite movements and Markovian rainfall changes, while the large-timescale satellite band switching is formulated as an integer programming problem. A hierarchical algorithm based on the deep reinforcement learning (DRL) approach and the greedy method is proposed to make efficient decisions. Extensive simulations are conducted to validate the effectiveness and advantages of our proposed scheme.**

***Index Terms*— LEO satellite, uplink transmission, multiband antenna, resource allocation, deep reinforcement learning.**

I. Introduction

THE rapid growth of the mobile Internet and the Internet of Things (IoT) has triggered an ever-increasing demand for massive connections and high-data-rate transmission [1]. However, the high terrestrial network deployment cost with limited backhaul transmission capacity poses challenges on providing high-data-rate communication services, especially in remote areas [2], [3]. Fortunately, by launching thousands of low earth orbit (LEO) satellites, ultra-dense LEO satellite networks can be built up to support high-capacity backhaul, seamless coverage, and flexible network access [4]. Aiming to improve the network data rate, many studies have been conducted in ultra-dense LEO satellite networks from resource allocation, handover management, and data offloading. Abdelsadekn *et al.* present a joint optimization framework for power allocation and handover management to optimize network throughput and handover rate [5]. Li *et al.* propose a user-centric handover scheme to leverage the satellite storage capacity to improve the performance of user access services [6]. Deng *et al.* design a game-based data offloading and spectrum allocation scheme to maximize the utility of terrestrial and satellite operators [7].

Most existing works consider satellites of different altitudes working over the same band, which may cause severe co-channel interference among satellites in an ultra-dense network [4]. Moreover, as Ka/Ku-band resources become scarce in recent years, higher bands (such as Q-band) are exploited for more bandwidth, which may cause higher rain attenuation. Therefore, the satellite bands need to be dynamically adjusted to balance the available spectrum resources, co-channel interference, and rain attenuation, thereby optimizing the overall data rate. Fortunately, this operation can be realized by ultra-dense LEO satellite networks with multiband antennas [8].

Although multiband ultra-dense satellite networks are promising in enhancing network data rate, challenges also exist in designing the transmission strategy. On the one hand, the movement of LEO satellites causes variations in the number of ground users, elevation angles, and transmission distances, which affects the interference. In addition, the changing rainfall has a huge impact on channel attenuation. Therefore, an efficient transmission resource allocation strategy is needed to adapt to the network dynamics. On the other hand, switching among different bands brings about additional power consumption and signaling overhead, which needs to be executed in a larger timescale. As these two decisions (i.e., resource allocation and band switching) are coupled, this letter proposes a two-timescale joint optimization framework to maximize the overall data rate while reducing the band switching cost in multiband ultra-dense satellite networks. We describe the small-timescale resource allocation problem as a Markov decision process (MDP) and the large-timescale band switching problem as an integer programming. A hierarchical algorithm based on the deep reinforcement learning (DRL) approach and the greedy method is proposed to solve the two-timescale joint optimization problem. The overall performance is maximized through continuous interaction between the problems of the two timescales.

II. System Model

Consider a time-slotted system where the slot index is denoted by t $(t = 0, 1, \ldots)$. As shown in Fig. 1, in the multiband ultra-dense LEO satellite network, a set of LEO satellites denoted by $\mathcal{V} = \{1, 2, \ldots, V\}$ fly over an area of interest to serve a set of ground users denoted by $\mathcal{N}(t) = \{1, 2, \ldots, N(t)\}$. $N(t)$ is the number of ground users served by satellites in time slot t and changes with the satellite movements. Denote the set of available band types as $\mathcal{K} = \{1, 2, \ldots, K\}$, and $\gamma_v(t) = k$ $(k \in \mathcal{K})$ means that satellite v operates in band k in time slot t. The bandwidth of each band is W_k, and is divided as a set of subchannels $\mathcal{M} = \{1, 2, \ldots, M\}$. The location of satellite and ground user is

Manuscript received February 20, 2022; accepted March 16, 2022. Date of publication March 21, 2022; date of current version June 10, 2022. This work was supported by the National Natural Science Foundation of China (No. 62171064) and the Beijing Municipal Science and Technology Project (No. Z211100004421019). The associate editor coordinating the review of this letter and approving it for publication was L. Mucchi. *(Corresponding author: Jiang Liu.)*

Tianjiao Chen, Jiang Liu, Qinqin Tang, Tao Huang, and Yunjie Liu are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with the Purple Mountain Laboratories, Nanjing 211111, China (e-mail: 2013210074@bupt.edu.cn; liujiang@bupt.edu.cn; qqtang@bupt.edu.cn; htao@bupt.edu.cn; liuyj@chinaunicom.cn).

Qiang Ye is with the Department of Computer Science, Memorial University of Newfoundland, St. John's, NL A1B 3X5, Canada (e-mail: qiangy@mun.ca).

Weiting Zhang is with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: wtzhang@bjtu.edu.cn).

Digital Object Identifier 10.1109/LCOMM.2022.3160839

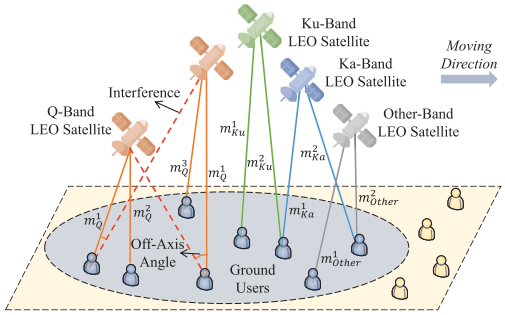

Fig. 1. Illustration of multiband ultra-dense LEO satellite networks.

denoted by $\boldsymbol{l}_v(t)$ and $\boldsymbol{l}_n(t)$ in a three-dimensional Euclidean coordinate. Equipped with multiple independent antenna apertures, each ground user can transmit data to multiple satellites simultaneously [4]. A controller is deployed on a geostationary earth orbit (GEO) satellite to make decisions centrally.

The uplink channel gain between ground users n and satellite v over subchannel m in time slot t is given by [9]

$$|h_{n,v,m}(t)|^2 = \frac{g_{n,v,m}(t)g_{v,m}(t)}{u^f_{n,v,m}(t)u^r_{n,v,m}(t)} \tag{1}$$

where $g_{n,v,m}(t)$ and $g_{v,m}(t)$ are the ground user transmitting antenna gain and the satellite receiving antenna gain. $u^f_{n,v,m}(t)$ is the free space path loss and $u^r_{n,v,m}(t)$ is rain attenuation. $g_{n,v,m}(t)$ and $g_{v,m}(t)$ is given by [10]

$$\begin{aligned} g_{n,v,m}(t) &= \phi_n \left(\frac{\pi f_{\gamma_v(t),m}\Omega_n}{c}\right)^2, \\ g_{v,m}(t) &= \phi_v \left(\frac{\pi f_{\gamma_v(t),m}\Omega_v}{c}\right)^2 \end{aligned} \tag{2}$$

where ϕ is the efficiency of the antenna, Ω is diameter of the antenna, c is the velocity of light and $f_{\gamma_v(t),m}$ is the communications center frequency of uplinks.

The free space path loss $u^f_{n,v,m}(t)$ is given by

$$u^f_{n,v,m}(t) = \left(\frac{4\pi\Gamma_{n,v}(t)}{\lambda_{\gamma_v(t),m}}\right)^2, \quad \lambda_{\gamma_v(t),m} = c/f_{\gamma_v(t),m} \tag{3}$$

where $\Gamma_{n,v}(t) = \|\boldsymbol{l}_v(t) - \boldsymbol{l}_n(t)\|_2$ is the distance between LEO satellite v and ground user n.

The rain attenuation $u^r_{n,v,m}(t)$ is mainly affected by frequency, elevation angle, altitude above sea level, and rainfall intensity, given by

$$u^r_{n,v,m}(t) = \Gamma^r_{n,v}(t)\rho_{\gamma_v(t),m} \cdot \varrho_n(t)^{\eta_{\gamma_v(t),m}} \tag{4}$$

where $\Gamma^r_{n,v}(t)$ is the effective path length of the wave in the rain, which is a function of $\boldsymbol{l}_v(t)$ and $\boldsymbol{l}_n(t)$ [9]. The rain intensity is $\varrho_n(t)$. ρ and η are derived from the frequency-dependent coefficients [9].

According to the Shannon formula, the achievable uplink data rate $R_{n,v,m}(t)$ (in bps) from ground user n to satellite v over subchannel m in time slot t is

$$R_{n,v,m}(t) = \frac{W_{\gamma_v(t)}}{M}\log_2 \times \left(1 + \frac{p_{n,v,m}(t)b_{n,v,m}(t)|h_{n,v,m}(t)|^2}{I_{n,v,m}(t) + \sigma^2}\right). \tag{5}$$

Here $b_{n,v,m}(t) = 1$ indicates that ground user n is associated with satellite v over subchannel m and $b_{n,v,m}(t) = 0$ otherwise. $W_{\gamma_v(t)}$ is the total bandwidth of satellite v, $p_{n,v,m}(t)$ is the transmission power, and N_u is the noise.

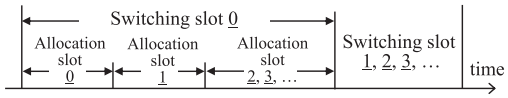

Fig. 2. Illustration of the two-timescale decision.

$I_{n,v,m}(t)$ is the co-channel interference between ground users n and satellite v over subchannel m and is given by

$$I_{n,v,m}(t) = \sum_{n'\neq n}\sum_{v'\neq v}\chi_{v,v'}(t)b_{n',v',m}(t) \cdot p_{n',v',m}(t)\frac{g(\varphi_{v,n',v'})(t)g_{v,m}(t)}{u^f_{n',v,m}(t)u^r_{n',v,m}(t)}\Big) \tag{6}$$

where $\chi_{v,v'}(t) = \mathbf{1}(\gamma_v(t) = \gamma_{v'}(t))$ and $\mathbf{1}(\cdot)$ is the indicator function. $g(\varphi_{v,n',v'})(t)$ is the off-axis antenna gain and is a function of off-axis angle $\varphi_{v,n',v'}$ [11].

III. Joint Transmission Resource Allocation and Band Switching Framework

A. Problem Statement

In this section, we formulate the joint transmission resource allocation and band switching optimization problem. As shown in Fig. 2, due to the asynchronous occurrence of resource allocation and band switching, the joint optimization problem is formulated as a two-timescale framework. Ground users choose the satellites to access with the reasonable power and subchannel allocation in a small timescale (i.e., allocation slots) to maximize the data rate. Satellites perform band switching on a larger timescale (i.e., switching slots) to maximize the overall data rate while reducing the switching cost. Due to the correlation between the two timescale problems, a hierarchical optimization framework is established.

B. Small-Timescale Transmission Resource Allocation

In the small-timescale transmission resource allocation, our objective is to maximize the total data rate of ground users. Considering continuous LEO satellite movements and Markovian rainfall changes, we describe the transmission resource allocation problem as an MDP. The MDP can be presented as a tuple $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} >$, where $\mathcal{S}$ denotes the set of states, $\mathcal{A}$ denotes the set of actions, $\mathcal{P}$ is a state transition probability, and $\mathcal{R}$ is the reward function.

- The network state includes the locations of all ground users and satellites and the rain intensities of all ground users, and is defined as $\boldsymbol{s}_t = (\boldsymbol{l}_n(t), \boldsymbol{l}_v(t), \varrho_n(t)), \forall n \in \mathcal{N}(t), v \in \mathcal{V}$.
- The system action includes the subchannel and the power allocation, and is defined as $\boldsymbol{a}_t = (b_{n,v,m}(t), p_{n,v,m}(t)), \forall n \in \mathcal{N}(t), v \in \mathcal{V}, m \in \mathcal{M}$.
- The immediate reward is the total data rate of all ground users, which is $r_t = \sum_{n\in\mathcal{N}}\sum_{v\in\mathcal{V}}\sum_{m\in\mathcal{M}} R_{n,v,m}(t)$.

The problem objective is to determine a set of stationary policies denoted by $\boldsymbol{\pi}$ to maximize reward. Then, we formulate the small-timescale transmission resource allocation optimization problem as follows:

$$\text{(P1):} \quad \max_{\boldsymbol{\pi}} \quad \mathcal{R}(\boldsymbol{\pi}) = \lim\sup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} r_t|\boldsymbol{\pi}\right] \tag{7}$$

$$\text{s.t. } b_{n,v,m}(t) \in \{0,1\}, \quad \forall n \in \mathcal{N}(t), \; v \in \mathcal{V}, \; m \in \mathcal{M} \tag{8}$$

$$\sum_{n\in\mathcal{N}} b_{n,v,m}(t) \leqslant 1, \quad \forall v \in \mathcal{V}, \; m \in \mathcal{M} \tag{9}$$

$$\sum_{v\in\mathcal{V}}\sum_{m\in\mathcal{M}} b_{n,v,m}(t) \leqslant V^r, \quad \forall n \in \mathcal{N}(t) \tag{10}$$

$$\sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}} p_{v,m}(t) b_{n,v,m}(t) \leqslant p_n^{\max}, \quad \forall v \in \mathcal{V} \tag{11}$$

where $\mathbb{E}[\cdot]$ is the expected reward. Constraint (8) and (9) indicate that each subchannel can be allocated to one ground user. Constraint (10) means that each ground user is associated to at most V^r subchannels simultaneously. $p_n^{\max}$ is the available power of ground user n and (11) is the power constraint.

C. Large-Timescale Band Switching

Considering the rainfall changes and the co-channel interference, the band of satellites needs to be adjusted to optimize the data rate while reducing the cost of band switching. In each switching slot, the band of satellites is fixed as $\boldsymbol{\gamma} = \{\gamma_v | v \in \mathcal{V}\}$ to indicate the band type of satellites, where $\gamma_v = k$, $k \in \mathcal{K}$. The switching cost is related to the band at the previous moment. Given the band type of the previous switching slot as $\boldsymbol{\gamma}'$, the large-timescale band switching problem is presented as an integer nonlinear programming:

$$\text{(P2):} \quad \max_{\boldsymbol{\gamma}} \mathcal{R}(\boldsymbol{\gamma}, \boldsymbol{\pi}) - \kappa \mathcal{C}(\boldsymbol{\gamma}', \boldsymbol{\gamma}) \tag{12}$$

$$\text{s.t. } \gamma_v \in \mathcal{K}, \forall v \in \mathcal{V} \tag{13}$$

Here, κ is a wight parameter. $\mathcal{R}(\boldsymbol{\gamma}, \boldsymbol{\pi})$ is the long-term data rate given $\boldsymbol{\gamma}$ and the small-timescale policy $\boldsymbol{\pi}$. $\mathcal{C}(\boldsymbol{\gamma}', \boldsymbol{\gamma}) = \sum_{v\in\mathcal{V}} \mathbf{1}(\gamma_v \neq \gamma_v')$ is the switching cost.

D. Two-Timescale Joint Optimization Framework

From the two timescale problem formulations, the resource allocation policy obtained from (P1) depends on how the band of satellites is selected. If the satellite band selection is inappropriate, the total data rate may decrease. Therefore, the two-timescale problems (P1) and (P2) should be solved simultaneously to achieve the larger data rate while reducing the band switching cost. Here, we propose a joint transmission resource allocation and band switching optimization framework to determine the efficient satellite band switching policy, γ_v^*, and the efficient subchannel allocation $b_{n,v,m}^*$, with the allocated power $p_{n,v,m}^*$.

IV. Learning-Assisted Hierarchical Approach

A. DRL for Transmission Resource Allocation

To solve the small-timescale optimization problem (P1), a DRL algorithm based on soft actor-critic (SAC) is designed.

1) SAC Framework: In SAC framework, the learning agent deployed on the GEO satellite has been split into two separate entities: the actor (policy) and the critic (value function) [12]. The goal of SAC is to find a resource allocation policy $\pi(\boldsymbol{a}|\boldsymbol{s})$ to maximize the expected reward. An entropy $\mathcal{H}(\pi(\boldsymbol{a}|\boldsymbol{s}))$ is added to the reward to ensure continually exploration. To evaluate policy π, the soft Q-value function is estimated as:

$$Q^{\pi}(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t + \alpha \sum_{t=1}^{\infty} \gamma^t H(\pi(\cdot|\boldsymbol{s}_t)) \,|\, \boldsymbol{s}_0 = \boldsymbol{s}, \boldsymbol{a}_0 = \boldsymbol{a}\right] \tag{14}$$

Here, $\mathcal{H}(\pi(\cdot|\boldsymbol{s})) = \mathbb{E}_{\boldsymbol{a}}[-\log \pi(\boldsymbol{a}|\boldsymbol{s})]$; α is the temperature parameter and γ is the discount factor. Through fully connected DNNs, the soft Q-value function and the policy are then can be parameterized as $Q_{\boldsymbol{\theta}}(\boldsymbol{s}, \boldsymbol{a})$ and $\pi_{\boldsymbol{\psi}}(\boldsymbol{a}|\boldsymbol{s})$ with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, respectively.

2) Algorithm Design: Based on the above framework, we design a SAC-based algorithm consisting of multiple training episodes. An episode contains multiple steps. In this letter, an episode corresponds to a switching slot, and a step corresponds to an allocation slot.

- Initialization: The parameters of soft Q-value functions/policies and the replay memory $\mathcal{M}_r$ are initialized.

- Experience sampling: In each step, the learning agent observes the network state $\boldsymbol{s}_t$. According to state $\boldsymbol{s}_t$ and policy $\pi_{\boldsymbol{\psi}}(\boldsymbol{a}_t|\boldsymbol{s}_t)$, the learning agent can generate action $\boldsymbol{a}_t$. Afterwards, the feedback reward r_t and the next state $\boldsymbol{s}_{t+1}$ can be obtained. The experience $\{\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1}\}$ is then stored into the experience replay memory $\mathcal{M}_r$.

- Parameter updating: The actor and the critic are updated by randomly sampling a mini-batch $\mathcal{M}_b$ of tuples $\{\boldsymbol{s}_t^m, \boldsymbol{a}_t^m, r_t^m, \boldsymbol{s}_{t+1}^m\}$ from $\mathcal{M}_r$.

a) Critic updating: Two separate critic networks are designed to avoid the overestimation. The parameters $\boldsymbol{\theta}_d$, $\forall d \in \{1, 2\}$ of the evaluation critic networks are updated via minimizing the loss function, which is given by

$$\mathcal{L}(\boldsymbol{\theta}_d) = \mathbb{E}_{\mathcal{M}_b}\left[\frac{1}{2}\left(Q_{\boldsymbol{\theta}_d}(\boldsymbol{s}_t, \boldsymbol{a}_t) - y_t\right)^2\right] \tag{15}$$

where y_t is target soft value with target parameter $\bar{\boldsymbol{\theta}}_d$, $\forall d \in \{1, 2\}$. It can be calculated as:

$$y_t = r_t + \gamma\Big(\min_{d=1,2} Q_{\bar{\boldsymbol{\theta}}_d}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) - \alpha \log\left(\pi_{\boldsymbol{\psi}}(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1})\right)\Big) \tag{16}$$

Then, the parameters $\boldsymbol{\theta}_d$, $\forall d \in \{1, 2\}$ are updated by computing the gradients of $\mathcal{L}(\boldsymbol{\theta}_d)$. Moreover, denote $\tau \in (0, 1)$ as the update factor, the target parameters $\bar{\boldsymbol{\theta}}_d$ are updated as

$$\bar{\boldsymbol{\theta}}_d = \tau \boldsymbol{\theta}_d + (1 - \tau)\bar{\boldsymbol{\theta}}_d, \forall d = 1, 2 \tag{17}$$

b) Actor updating: The policy parameters of the actor can be learned by minimizing the expected KL-divergence:

$$J(\boldsymbol{\psi}) = \mathbb{E}_{\mathcal{M}_b}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi_{\boldsymbol{\psi}}(\cdot|\boldsymbol{s}_t) \,\middle\|\, \frac{\exp\left(\min_{d=1,2} Q_{\boldsymbol{\theta}_d}(\boldsymbol{s}_t, \cdot)\right)}{Z_{\boldsymbol{\theta}_d}(\boldsymbol{s}_t)}\right)\right] \tag{18}$$

where $Z_{\boldsymbol{\theta}_d}(\boldsymbol{s}_t)$ is an intractable partition function that does not contribute to the new policy's gradient. Therefore, we further transform the above KL-divergence as:

$$J(\boldsymbol{\psi}) = -\mathbb{E}_{\mathcal{M}_b}\Big[\mathop{\mathbb{E}}_{\boldsymbol{a}_{t+1}\sim\pi_{\boldsymbol{\psi}}}\Big[\min_{d=1,2} Q_{\bar{\boldsymbol{\theta}}_d}(\boldsymbol{s}_{t+1}^m, \boldsymbol{a}_{t+1}^m) - \alpha \log\left(\pi_{\boldsymbol{\psi}}(\boldsymbol{a}_{t+1}^m|\boldsymbol{s}_{t+1})\right)\Big]\Big] \tag{19}$$

Then, the parameters $\boldsymbol{\psi}$ of policy is updated by computing the gradient of $J(\boldsymbol{\psi})$.

c. Temperature updating: The temperature parameter can be updated automatically by calculating to the gradient of the following objective:

$$J(\alpha) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[-\alpha \log\left(\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)\right) - \alpha \bar{H}\right] \tag{20}$$

where $\bar{H}$ is the value of target entropy.

Algorithm 1 Learning-Assisted Hierarchical Solution for Joint Optimization

Input: $\boldsymbol{\gamma}'$; $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\gamma}'$; $j = 0$.
Output: A set of efficient solutions for resource allocation policy $\boldsymbol{\pi}^*$, satellite band switching policy $\boldsymbol{\gamma}^*$.

repeat
  $j \leftarrow j + 1$;
  Update the operating band of satellites based on $\boldsymbol{\gamma}^{(j-1)}$;
  With $\boldsymbol{\gamma}^{(j-1)}$, run **Sub-algorithm 1** to determine the stationary resource allocation policy $\pi_{\boldsymbol{\psi}}^{(j)}$;
  Given $\boldsymbol{\pi}^{(j)} = \pi_{\boldsymbol{\psi}}^{(j)}$, run **Sub-algorithm 2** to obtain the band switching policy $\boldsymbol{\gamma}^{(j)}$ in iteration j ;
until $\boldsymbol{\pi}^{(j)} = \boldsymbol{\pi}^{(j-1)}$;
Sub-algorithm 1 `SAC-based Resource Allocation`**:**
  Receive $\boldsymbol{\gamma}$ and initialize SAC parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$;
  for $episode = 1, 2, \ldots$ **do**
    Initialize environment and receive initial state $\boldsymbol{s}_0$;
    for $t = 1, 2, \ldots, T$ **do**
      Given $\boldsymbol{\gamma}$, observe the current state: $\boldsymbol{s}_t = (\boldsymbol{l}_n(t), \boldsymbol{l}_v(t), \varrho_n(t)), \forall n, v$;
      Select $\boldsymbol{a}_t$ with policy $\pi_{\boldsymbol{\psi}}(\boldsymbol{a}_t|\boldsymbol{s}_t)$;
      Obtain the reward r_t and the next state $\boldsymbol{s}_{t+1}$;
      Store the tuple $\{\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1}\}$ into $\mathcal{M}_r$;
      Randomly sample a mini-batch of tuples from $\mathcal{M}_r$;
      Compute the target Q-value y_t by Eq. (16);
      Update soft Q-value function parameters $\boldsymbol{\theta}_d$ by minimizing the loss function in Eq. (15);
      Update parameter $\boldsymbol{\psi}$ via the gradient in Eq. (19);
      Update temperature parameter α by computing the gradient defined in Eq. (20);
      Update target networks with Eq. (17);
    end
  end
return $\pi_{\boldsymbol{\psi}}$
Sub-algorithm 2 `Greedy-based Band Switching`**:**
  Receive $\boldsymbol{\pi}$ and initialize $\mathcal{W}_v = \{\omega_v^1, \ldots, \omega_v^K\}, \forall v \in \mathcal{V}$;
  $\mathcal{W} = \cup_v \mathcal{W}_v$; $\Psi = \emptyset$;
  while $\mathcal{W} \neq \emptyset$ **do**
    Let $\omega_v^k = \arg\max_{\omega \in \mathcal{W}} f(\omega|\Psi)$;
    $\Psi = \Psi \cup \{\omega_v^k\}$ and $\mathcal{W} = \mathcal{W}/\mathcal{W}_v$;
  end
  $\boldsymbol{\gamma} = g(\boldsymbol{\gamma}', \Psi)$;
return $\boldsymbol{\gamma}$

B. Hierarchical Solution for Joint Optimization

In the proposed two-timescale joint framework, problems (P1) and (P2) are solved iteratively for the transmission resource allocation policy and the band switching policy. Specifically, (P2) is solved by the greedy algorithm and then the joint efficient solutions are iteratively obtained.

A brute-force attack algorithm can be adopted to solve problem (P2), and its computational complexity is $\mathcal{O}(K^V)$. Thus, to reduce the computational complexity, a greedy-based algorithm is employed in our work. We first define the marginal utility as follows.

Definition 1. (Marginal Utility) Given a valuation f on a finite set Ω and a set $X \subseteq \Omega$, and then the marginal utility of a set of $Y \subseteq \Omega - X$ is given by

$$f(Y|X) = f(X \cup Y) - f(X). \tag{21}$$

For problem (P2), we define a ground set $\mathcal{W}$. By using the element ω_v^k to indicate that the satellite v switches to band k, $\forall k \in \mathcal{K}$, the ground set is written as $\mathcal{W} = \{\omega_1^1, \ldots, \omega_V^1, \ldots, \omega_1^K, \ldots, \omega_V^K\}$. The ground set is partitioned into V disjoint sets, $\mathcal{W}_1, \ldots, \mathcal{W}_v, \ldots, \mathcal{W}_V$, where $W_v = \{\omega_v^1, \ldots, \omega_v^K\}$ is the set of band that satellite v might be switch to. Denote $\Psi \subseteq \mathcal{W}$ as a solution set of (P2). Then, we can rewrite (12) as $\mathcal{G}(\Psi) = \mathcal{R}(g(\boldsymbol{\gamma}', \Psi), \boldsymbol{\pi}) - \mathcal{C}(\boldsymbol{\gamma}', g(\boldsymbol{\gamma}', \Psi))$, where $g(\boldsymbol{\gamma}', \Psi)$ is a mapping function. With such a function, we can obtain $\boldsymbol{\gamma} = \{\gamma_v = \gamma_v', \forall v \in \mathcal{V}; \gamma_v = k, \forall \omega_v^k \in \Psi\}$.

Based on the theories above, we adopt a greedy-based approach for the large-timescale problem. We first initialize $\Psi = \Psi'$, where Ψ' is a solution set mapped via $\boldsymbol{\gamma}'$. An element ${\omega'}_v^k \in \Upsilon'$ if $\gamma_v' = k$. In each iteration round, the element with the largest marginal value $f(\omega_v^k|\Psi) = \mathcal{G}(\Psi \cup \omega_v^k) - \mathcal{G}(\Psi)$ is selected from the ground set $\mathcal{W}$. When a satellite v switches to a new band k, the related element is removed from $\mathcal{W}$. Through greedy operation, efficient band switching is realized.

As shown in *Algorithm 1*, we propose a learning-assisted hierarchical solution to solving (P1) and (P2) iteratively. (P1) is solved for resource allocation policy $\boldsymbol{\pi}$ which is then fed into (P2) as input to obtain the band switching policy $\boldsymbol{\gamma}$. For given $\boldsymbol{\pi}$, the band switching is adjusted in proportion to the power and channel allocation. Therefore, even if the pre-configured satellite band setting is inappropriate, $\boldsymbol{\gamma}$ is gradually optimized. The convergence of SAC-based algorithm has been proved by [13]. A greedy algorithm can obtain a local optimal solution that approximates the global optimal solution [14]. Then, the efficient resource allocation policy and band switching policy can be jointly obtained. The computational complexity of SAC comes down to the parameters update of two DNNs (i.e., the actor and the critic), which is $\mathcal{O}(wBTZ)$ [13]. Z is mini-batch size, and w, B, T are the number of neurons, episodes and steps, respectively. For the greedy-based algorithm, the computational complexity is $\mathcal{O}(VK)$. Assuming that the convergence is reached after J iterations, the total computational complexity of *Algorithm 1* is calculated as $\mathcal{O}(J \cdot (wBTZ + VK))$.

V. Numerical Results and Discussion

Simulations are conducted in an ultra-dense LEO satellite scenario consisting of 5 LEO satellites. The ground users are randomly distributed on the ground. The heights of satellites vary from 580 km to 620 km, and the minimum elevation angle is 40°. Each switching slot has 100 allocation slots. In one allocation slot, an average of 6 users with two antennas are under coverage. Each time slot lasts 4 s. Q-, Ka- and Ku-band are considered, and the bandwidth of these three bands are 1.3, 1 and 0.7 MHz, respectively. The noise density is -174 dBm/Hz. The vertical equivalent path length of rain is 4.5 km given by P.1151 Rec. ITU-R. The efficiency and diameter of the satellite antenna are 0.6 and 2 m, respectively [10]. For the ground users such as vehicle terminals, the efficiency and diameter of the antenna are 0.55 and 1 m, respectively. The power of each ground user is 10 W. We set 4 levels of rain intensity $[0, 6, 12, 18]$ mm/h. As the rain intensity transition probability parameter (TPP) increases, the probability of heavy rain will decrease. For the SAC-based algorithm, the replay memory size is 2×10^4, and the mini-batch size is 128. The discount factor and the learning rate are set to 0.85 and 0.0005.

Fig. 3 shows the convergence performance of the proposed scheme. First, we compare it with the deep deterministic policy

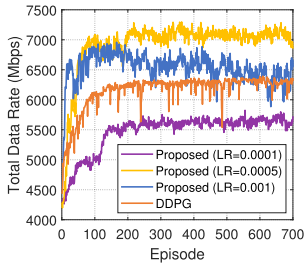

Fig. 3. Multiband performance.

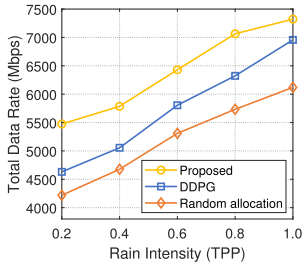

Fig. 4. Small-timescale allocation.

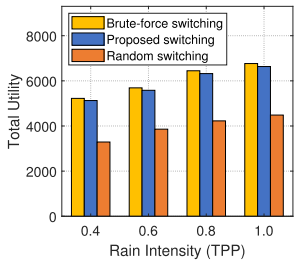

Fig. 5. Large-timescale switching.

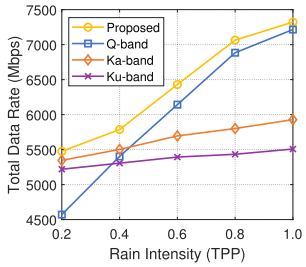

Fig. 6. Multiband performance.

gradient (DDPG)-based scheme. The proposed scheme has a faster learning convergence speed and a better reward than the DDPG-based scheme. As the learning rate increases, the algorithm converges faster. However, fast convergence may cause the agent to fall into the local optimum. Therefore, we need to select an appropriate learning rate.

Due to the difference of Q-, Ka- and Ku-band affected by rain, we compared the performance of four schemes under different rainfall. Fig. 4 compares the proposed small-timescale resource allocation algorithm with the DDPG-based and the random allocation scheme. With the decrease of rain intensity TPP, the total data rate increases, and the performance of the proposed SAC-based scheme is the best. Fig. 5 compares the proposed large-timescale band switching algorithm with brute-force and random switching scheme. With the decrease of rain intensity, the total utility increases. The performance of the proposed greedy scheme is greater than the random scheme and is almost the same as the brute-force scheme.

In Fig. 6, we compare the performance of the proposed scheme with the fixed band schemes. The Ku-band scheme's bandwidth is 0.9 MHz. When there is no rain, Q-band scheme performs better in data rate due to its highest available bandwidth, and Ku-band is the lowest. When the rainfall increases, the performance of Q-band decreases sharply compared with Ka- and Ku-band. The proposed scheme has the largest data rate under different rain intensities, because the use of multiband reduces co-channel interference, and the dynamic band switching is adapted to different rain changes.

VI. Conclusion

This letter proposes an efficient uplink transmission scheme over ultra-dense LEO satellite networks with multiband antennas. We formulate the joint transmission resource allocation and band switching as a two-timescale framework composed of an MDP and an integer programming problem to maximize the overall data rate while reducing the band switching cost. A hierarchical algorithm based on the DRL approach and greedy method is proposed to make efficient decisions. Numerical results show the advantages of our proposed scheme. Running DRLs centrally on one GEO satellite may become a bottleneck when dealing with large-scale problems, hence we will study distributed transmission schemes in the future.

References

[1] Y. Zhang, Z. Xiong, D. Niyato, P. Wang, and Z. Han, "Information trading in Internet of Things for smart cities: A market-oriented analysis," *IEEE Netw.*, vol. 34, no. 1, pp. 122–129, Jan. 2020.

[2] H. Peng, H. Wu, and X. S. Shen, "Edge intelligence for multi-dimensional resource management in aerial-assisted vehicular networks," *IEEE Wireless Commun.*, vol. 28, no. 5, pp. 59–65, Oct. 2021.

[3] C. Zhou *et al.*, "Deep reinforcement learning for delay-oriented IoT task scheduling in SAGIN," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 911–925, Feb. 2021.

[4] B. Di, H. Zhang, L. Song, Y. Li, and G. Y. Li, "Ultra-dense LEO: Integrating terrestrial-satellite networks into 5G and beyond for data offloading," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 47–62, Jan. 2019.

[5] M. Y. Abdelsadek, H. Yanikomeroglu, and G. K. Kurt, "Future ultra-dense LEO satellite networks: A cell-free massive MIMO approach," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 1–6.

[6] J. Li, K. Xue, J. Liu, and Y. Zhang, "A user-centric handover scheme for ultra-dense LEO satellite networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1904–1908, Nov. 2020.

[7] R. Deng, B. Di, S. Chen, S. Sun, and L. Song, "Ultra-dense LEO satellite offloading for terrestrial networks: How much to pay the satellite operator?" *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6240–6254, Oct. 2020.

[8] S. D. Targonski, "A multiband antenna for satellite communications on the move," *IEEE Trans. Antennas Propag.*, vol. 54, no. 10, pp. 2862–2868, Oct. 2006.

[9] D. Zhou, M. Sheng, R. Liu, Y. Wang, and J. Li, "Channel-aware mission scheduling in broadband data relay satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 5, pp. 1052–1064, May 2018.

[10] Z. S. G. Maral and M. Bousquet, *Satellite Communications Systems: Systems*, 6th ed. Hoboken, NJ, USA: Wiley, 2020.

[11] *Radio Regulations (Edition 2012, Part 2)*, ITU, Geneva, Switzerland, 2012.

[12] W. Zhang *et al.*, "Optimizing federated learning in distributed industrial IoT: A multi-agent approach," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3688–3703, Dec. 2021.

[13] C. Sun, X. Wu, X. Li, Q. Fan, J. Wen, and V. C. M. Leung, "Cooperative computation offloading for multi-access edge computing in 6G mobile networks via soft actor critic," *IEEE Trans. Netw. Sci. Eng.*, early access, Apr. 30, 2021, doi: 10.1109/TNSE.2021.3076795.

[14] S. Ren, T. Lin, W. An, Y. Li, Y. Zhang, and Z. Xu, "Collaborative EPC and RAN caching algorithms for LTE mobile networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.